Combining microarrays and genetic analysis

Alberts, Rudi; Fu, Jingyuan; Swertz, Morris A.; Lubbers, L. Alrik; Albers, Casper J.; Jansen, Ritsert C.

Published in:
Briefings in Bioinformatics

DOI:
10.1093/bib/6.2.135

Publication date:
2005

Combining microarrays and genetic analysis

Rudi Alberts, Jingyuan Fu*, Morris A. Swertz, L. Alrik Lubbers, Casper J. Albers and Ritsert C. Jansen*

Date received (in revised form): 11th April 2005

Rudi Alberts
is a PhD student at the Groningen Bioinformatics Centre of the University of Groningen, focusing on genetical genomics in the mouse.

Jingyuan Fu
is a PhD student at the Groningen Bioinformatics Centre of the University of Groningen, focusing on genetical genomics in *Arabidopsis*.

Morris A. Swertz
is a PhD student at the Groningen Bioinformatics Centre of the University of Groningen, focusing on information systems in the life sciences. He has developed an information system for microarray data by using advanced code generation technologies.

L. Alrik Lubbers
works at the Groningen Bioinformatics Centre of the University of Groningen as a software engineer, specialising on automatic code generation for information systems in the life sciences.

Casper J. Albers
works at the Groningen Bioinformatics Centre of the University of Groningen on the statistical analysis of bioinformatics data, in particular, microarray data.

Ritsert C. Jansen
leads the Groningen Bioinformatics Centre of the University of Groningen with research on statistics, information analysis and system development for 'omics' data.

Keywords: *microarray, gene expression, quantitative trait locus, QTL, two-colour microarray, oligonucleotide microarray*

Ritsert C. Jansen,
Groningen Bioinformatics Centre,
Kerklaan 30, 9751 NN,
Haren, the Netherlands

Tel: +31 50 3638089
Fax: +31 50 3633400
E-mail: r.c.jansen@rug.nl

Abstract

Gene expression can be studied at a genome-wide scale with the aid of modern microarray technologies. Expression profiling of tens to hundreds of individuals in a genetic population can reveal the consequences of genetic variation. In this paper it is argued that the design and analysis of such a study is not a matter of simply applying the existing and more-or-less standard computational tools for microarrays to a new type of experimental data. It is shown how to fully exploit the power of genetics through optimal experimental design and analysis for two major microarray technologies, cDNA two-colour arrays and Affymetrix short oligonucleotide arrays.

**The authors have contributed equally to this paper.*

INTRODUCTION

A genetic study of a quantitative (or complex) trait generally starts with a measurable trait and traces its variation down to the molecular level – variation in the expression levels and in the coding sequences of genes. Classical examples of quantitative traits are disease or yield. Gene expression levels – though biomolecular in nature – are also eligible examples of quantitative traits and can be analysed the same way as we can analyse classical quantitative traits: by quantitative trait locus (QTL) mapping. This strategy, when implemented for gene expression levels at a genome-wide scale, is coined 'genetical genomics' and will provide insight into the gene expression network as a whole and – more in general – improve our understanding of classical complex traits of interest.[1] This paper describes the following: (i) the general concept of genetical genomics, (ii) the application on two-colour cDNA microarrays, (iii) the application on short oligonucleotide arrays and (iv) current and upcoming information systems for genetical genomics. Finally, the reader is directed to the first and recent publications on still relatively small populations and prospects of genetical genomics in coming larger experiments are discussed.

CONCEPTS OF GENETICAL GENOMICS

For the moment, assume we have our favourite technology for profiling gene expression. What does it mean to apply this technology to a genetic population? How should we statistically analyse the individual gene expression levels? How do we make inferences across genes in order to (re-)construct gene regulatory networks? In this section the very basics of genetical genomics are described using hypothetical examples.[2] Applications on real data will follow in later sections of this paper.

A hypothetical example of how gene expression may vary between two (founder) parental strains and within their genetic offspring is shown in Figure 1. This example deals with recombinant inbred lines (RILs), a type of genetic population which has become very popular in genetic studies (Figure 1A). Starting from the F1 of two homozygous parents, say A and B, after eight or more successive generations of inbreeding, (almost) homozygous RIL offspring have been developed. The individuals of each

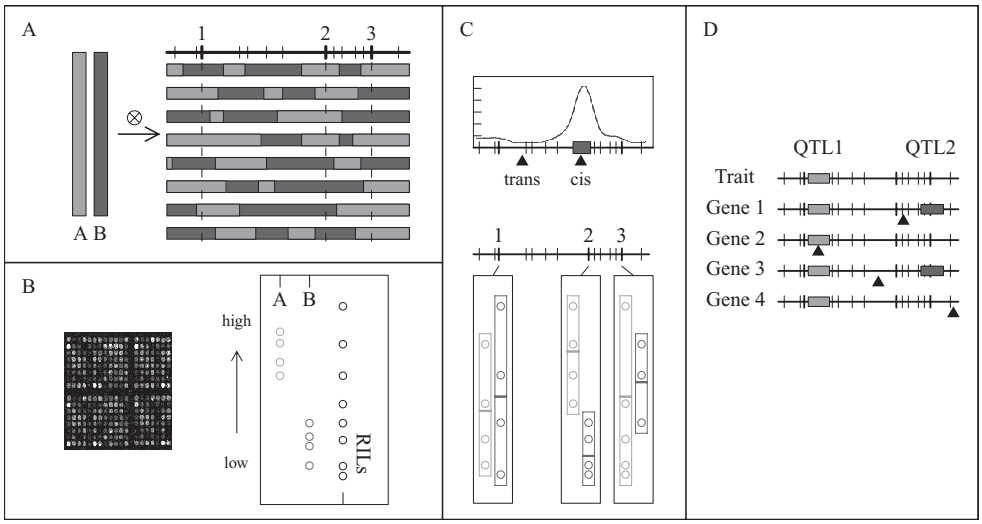

Figure 1: Genetic dissection of gene expression. (A) Genetics of recombinant inbred lines: a set of eight homozygous recombinant inbred lines (RILs) is generated from a cross between two homozygous parents (dark and light grey, respectively). The genomes of the RILs are a mosaic of the parental genomes; this mosaic can be viewed with the aid of molecular markers. (B) Expression profiling of parental strains and their RIL offspring. Example expression data for one gene, with four measurements for each parent and one measurement for each of the eight RILs. (C) Genetic dissection of gene expression. Allelic difference of expression is analysed at three example markers; means per group are indicated by a horizontal stripe. The significance can be plotted along the genome map, the peak indicates the position of a QTL underlying expression variation. (D) (Re-)construction of gene regulatory network. A phenotypic trait and the expression of four genes were analysed and mapped to a similar QTL region. The triangles indicate the locations of the genes themselves. Gene 2 is *cis*-acting and therefore this gene is a good candidate regulator of the trait and possibly also of the other genes. Gene 1 and 3 share two QTLs, which can be taken as evidence (but not as a proof) that these genes act in the same pathway

Genetic dissection of gene expression

generation are self crossed or sibling mated in order to generate the next generation. The genomes of the A×B RILs are a mosaic of the 'founder' genomes A and B, and this mosaic can be viewed with the aid of molecular markers (Figure 1A). Although generating RILs is an expensive and time-consuming process, RILs have many advantages over other segregating populations, such as F2 and backcross (BC) populations: they are homozygous and each strain is an eternal resource; each strain only needs to be genotyped once; and the denser breakpoints after many generations of inbreeding can increase mapping precision. An example of how a gene's expression levels can vary between the two parents (A and B) and across the RILs is shown in Figure 1B.

QTL: genetic loci that contribute to variability in quantitative (expression) traits

Figure 1C shows a typical output of QTL analysis. At three marker positions, the expression phenotype of the offspring carrying allele A is shown as well as that of the offspring carrying allele B. The expression difference between the two groups is most significant at marker 2 and the expression phenotype of this gene is said to map to marker 2 (see below for basic statistics). If QTL and gene co-localise, the gene is said to map to itself or to act 'in *cis*'. If QTL and gene do not co-localise, the gene is said to act 'in *trans*'. *Cis*-activity can be explained by altered functional motifs in the promoter region that will affect the initialisation of transcription; by altered sequence elements that will change the stability of the mRNA, for example, a sequence polymorphism in the 3′ untranslated

region (UTR); or by feedback control from a modified gene product. The mechanism of regulation 'in *trans*' is often complicated and wide-spread over different classes of genes.[3]

Narrowing down from QTL to gene

For each gene in turn, we can identify its *cis*- or *trans*-regulating elements. Genetic dissection of thousands of gene profiles gives thousands of pieces of the gene expression puzzle. How can we use this to narrow down from QTL to gene and to re-construct expression regulatory networks? Owing to the relatively low resolution of QTL mapping, each QTL region can still contain hundreds of candidate genes. One strategy to narrow a QTL region down to a more limited set of candidate genes is to proceed by making further crosses with a higher number of recombinants and a denser marker map. However, with all the expression data at hand, there is a powerful alternative based on co-localisation of QTLs. To illustrate this a hypothetical example is elaborated in which a trait and expression variation of multiple genes map to the same region (QTL1 in Figure 1D). To narrow down from QTL to causal gene, we search for *cis*-acting genes in this region. Only gene 2 is *cis*-acting and is therefore a good candidate for QTL1. Gene 2 may also be considered a candidate regulator of genes 1, 3 and 4. Co-localisation of expression variation at two or more QTLs provides stronger evidence for causal relationships between genes and can form the basis for gene network reconstruction; genes 1 and 3 both map to QTL1 and QTL2 and may be in the same pathway. The availability of QTL profiles for every gene allows the reconstruction of gene regulatory networks. The idea is that genes with similar QTL profiles could be in the same pathway, because they seem to have common regulators.

Gene network reconstruction

With the aid of molecular markers, the variation of gene expression can be mapped to QTL; we here outline the basic statistics. It is commonly assumed that the logarithm-transformed expression data follow a normal (Gaussian) distribution. For simplicity assume there is one causal gene (quantitative trait locus; QTL) underlying the variation. A general QTL model reads

$$y_{ij} = m + Q_{ij} + e_{ij}$$

where y_{ij} is the logarithm of the expression phenotype for the jth individual with the ith QTL genotype ($i = 1, 2$ corresponding to genotypes A and B, respectively), m is the overall mean of the expression level, Q_{ij} is the QTL effect, and the e_{ij} are the (environmental) errors, which are assumed to be independent of each other. To simplify the following formulas, let n_1 be the number of individuals with genotype 1 (A) and n_2 that of genotype 2 (B); let $y_{1\bullet}$, and $y_{2\bullet}$ be the marginal means of individuals with genotype A and B, respectively; let $y_{\bullet\bullet}$ be the overall mean. We do not observe the QTL genotype, but with a sufficiently dense marker map, we can search for a marker close to the causal gene (the QTL). A 'single marker' analysis can be adopted to find at which marker the differential expression between genotypes A and B is most significant. Statistics offers the well-known t-test to assess significance. The test statistic t is computed via

$$t = \frac{y_{1\bullet} - y_{2\bullet}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - y_{1\bullet})^2 + \sum_{j=2}^{n_2}(y_{2j} - y_{2\bullet})^2}{n_1 + n_2 - 2}$$

The t-test is known to be equal to the square root of the F-test in the basic analysis-of-variance (ANOVA) for QTL analysis (Table 1).

GENETICAL GENOMICS WITH TWO-COLOUR cDNA ARRAYS

In the two-colour cDNA microarray technology, two samples are labelled with different fluorescent dyes – usually, red

Table 1: Basic ANOVA table in QTL analysis

Source of variation	Sum of squares	d.f.	*F*-value
QTL	$SS_{between} = \left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{-1}(y_{1\bullet}-y_{2\bullet})^2$ $= n_1(y_{1\bullet}-y_{\bullet\bullet})^2 + n_2(y_{2\bullet}-y_{\bullet\bullet})^2$	1	$\frac{SS_{QTL}}{s^2}$
Error	$SS_{within} = (n_1+n_2-2)s^2$	n_1+n_2-2	
Total	$SS_{total} = \sum_{i=1}^{2}\sum_{j=1}^{n_j}(y_{ij}-y_{\bullet\bullet})^2$ $= SS_{QTL} + SS_{error}$	n_1+n_2-1	

Design of two-column micro-array experiments

(Cy5) and green (Cy3), and then mixed to co-hybridise on the microarray. Subsequently, the slide is scanned to obtain numerical intensities for each dye. Data of two-colour microarrays are generally pre-processed following the pipeline: background correction, dye-effect correction and normalisation within arrays and between arrays.[4] The next step of the statistical analysis usually relies on the log ratios of the two dyes, log(Cy5/Cy3), to assess the difference of expression levels. Analysing log ratios has the advantage that local spot errors (whether systematic or random in nature) cannot bias the results. Analysing intensities by using so-called mixed models[5] is a reasonable alternative provided that errors can be assumed to be randomly, independently and normally distributed. Two-colour microarrays pose new challenges in our genetical genomics set-up: how to optimally pair samples and how to model the expression QTL with ratios.[6] These two issues will be discussed below.

QTL analysis of ratios

The aim of the experiment is to detect genes that show differential expression between genotypes A and B. Two individuals can have different genotypes at a given gene (and the ratio is 'A/B' or 'B/A') or identical genotypes ('A/A' or 'B/B') – see Figure 2A for an example. The latter combination is uninformative and should be avoided. Thus, comparing genome 'AAABBAAAA. . .' with 'BBBAABBBB' provides informative ratios for more genes than comparing genomes 'AABAABBBB' and 'AAABBBBAA'. This recently proposed 'distant pair design' can significantly increase the efficiency of microarray and genotype resources when analysing log-ratios by using the above models or intensities by using the mixed-model.[6] Figure 2B shows the strategies of three competing experimental designs – common reference design, loop design and distant pair design. Figure 2C shows patterns of QTL co-localisation. All genes (*GI*, *LHY*, *CCA1*, *TOC1*, *ELF4*, *CO*, *FKF1*) are known to act in the long-day pathway and some of them clearly map to the same region at chromosome 1. *GI* is the *cis*-regulating gene in that region and it may be hypothesised that *GI* is the causal gene regulating the expression of the other genes.

At each spot we observe the intensities of two samples. Therefore, the expression QTL model should be modified. The (systematic or random) spot effect can be eliminated by taking ratios of co-hybridised samples. We have the informative ratios of type 'A/B' and 'B/A', and the uninformative ratios 'A/A' and 'B/B' ($i = 1,\ldots, 4$). Therefore, the QTL model on log-ratio scale reads:

$$r_{ij} = d + Q_{ij} + e_{ij}$$

where r_{ij} is the log-ratio of the gene expression levels, d is the overall

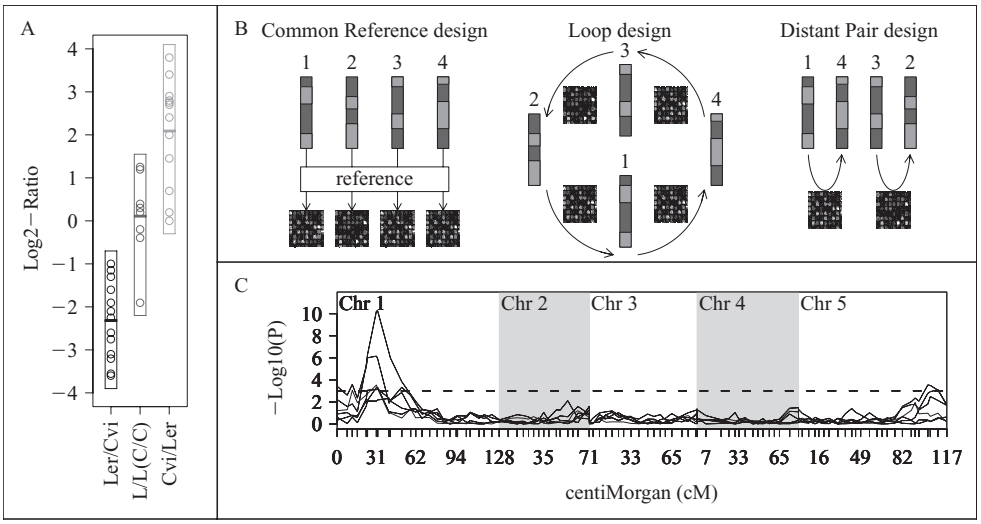

Figure 2: Design and analysis of two-colour microarrays. (A) Genetic dissection of gene expression. In a RIL population (derived from a cross between *Arabidopsis* ecotypes Ler and Cvi) the expression ratios were calculated. Marker HH.480C showed highly significant differential expression between the possible groups of genotype ratios 'Ler/Cvi', 'Cvi/Ler', 'Ler/Ler' and 'Cvi/Cvi'; means per group are indicated by a horizontal stripe. (B) Different designs are illustrated with four RILs. The common reference design hybridises individual samples to one and the same common reference. The loop design hybridises individual samples in a loop pattern. The distant pair design pair-wisely hybridises the individual samples that carry most different marker fingerprints. The arrows present the dye direction (green → red). (C) Genome-wide QTL profile of several flowering time genes using single marker analysis. Variation in expression of the FLC and other flowering time genes map to the same QTL. The gene GI is a *cis*-acting gene and a good candidate for the QTL

difference between the two different dyes, Q_{ij} is the QTL effect which takes the value q when $i = 1$, $-q$ when $i = 2$, and 0 when $i > 2$. We introduce the notations n_1, n_2, n_3 and n_4 for the number of measurements for A/B, B/A, A/A and B/B, respectively, $r_{1\bullet}$, $r_{2\bullet}$, $r_{3\bullet}$ and $r_{4\bullet}$ for the corresponding averages, n for the total number of measurements, and $r_{\bullet\bullet}$ for the grand mean. The corresponding ANOVA table can be found in Table 2. The statistical analysis is conducted at each marker along the genetic map, and we can therefore plot the F-value or P-value (or $-\log_{10}P$) as a measure for QTL significance.

Multiple QTL models

Ideally, when we test an interval for a QTL, we would like our test statistic to be independent of the effect of possible QTLs in other regions. The single QTL model can be extended to a multiple QTL model

$$r_{ij} = d + \sum_{k=1}^{K} Q_{ijk} + e_{ij}$$

where K is the number of QTLs. However, location and genotype of QTLs in other regions are not known exactly; fortunately it suffices to replace 'background' QTL by linked (adjacent) markers with known genotype. Such markers are included as cofactors in the model

$$r_{ij} = d + \sum_{m=1,\, m\neq k}^{K} M_{ijm} + Q_{ijk} + e_{ij}$$

where Q is the QTL of current interest and M are the effects of markers representing the background QTL.

Table 2: ANOVA table for ratios in two-colour microarrays

Source of variation	Sum of squares	d.f.	*F*-value
QTL effect	$SS_{QTL} = \sum_{i=1}^{4} n_i(r_{i\bullet} - r_{\bullet\bullet})^2$	1	$\frac{SS_{QTL}}{SS_{error}/(n-2)}$
Error	$SS_{error} = SS_{total} - SS_{QTL}$	$n-2$	
Total	$SS_{total} = \sum_{i=1}^{4}\sum_{j=1}^{n_j}(r_{i\bullet} - r_{\bullet\bullet})^2$	$n-1$	

GENETICAL GENOMICS WITH SHORT OLIGONUCLEOTIDE ARRAYS

In contrast to two-colour microarrays, short oligonucleotide arrays profile one sample per array. Consequently, oligonucleotide arrays provide only absolute mRNA abundance measurements per sample and no ratios between samples. A second difference is that short oligonucleotide arrays measure mRNA abundance by using multiple probes per gene, ie we have multiple observations per gene. This section discusses the consequences of these differences with respect to the dissection of gene expression variation and the statistical models used, and argues that standard methods for statistical analysis of this type of expression data do not extract all relevant information from the data.[7]

A commonly used oligonucleotide array is Affymetrix's GeneChip. Figure 3 shows example data on a recombinant inbred population of 30 mice derived from a cross between parental strains C57BL/6 (B6) and DBA/2 (D2). The samples were hybridised to Affymetrix MG-U74Av2 GeneChips containing 12,422 probe sets, each typically consisting of 16 probes. The mice were fingerprinted by using 779 molecular markers. The Robust Multi Array (RMA) background correction and normalisation procedures were applied. See Irizarry *et al.*[8] for an overview of normalisation methods for oligonucleotide arrays. RMA provides a method to summarise the multiple probe intensities into one expression value per probe set (gene). The use of this method is not advocated because relevant genetic information is thrown away by taking averages over probes.[7] Alternative approaches are described below. Figure 3A shows the relative location of the probes on the mRNA for an example probe set. Figure 3B shows the dissection of expression variation for each of the probes. The expression values are coloured dark and light grey depending on the allele type at the marker under study.

QTL analysis of probe data

The expression data on both probe and probe set level are modelled. Considering one probe at a time, for example probe 1 in Figure 3B, its expression variation can be dissected in a way similar to that in Figure 1C. The arrays in our experiment were obtained and processed in three batches. This unwanted source of extra variation is eliminated by taking batch as factor into the model. The ANOVA model for probe level analysis is

$$y_{ijk} = m + B_i + Q_j + e_{ijk}$$

where y_{ijk} is the logarithm of the probe intensity of the kth mouse ($k = 1, \ldots, n_{ij}$) from the ith batch ($i = 1, 2, 3$) with the jth QTL genotype ($j = 1, 2$, ie A and B, respectively). Here m is the overall mean, B_i is the batch effect, Q_j is the QTL effect, e_{ijk} is the error, n_{ij} are the number of mice from batch i with QTL genotype j and $\Sigma_i\Sigma_j n_{ij} = 30$. Some further useful notations are, in line with the previous section, $y_{\bullet\bullet\bullet}$ for the overall mean of all 30

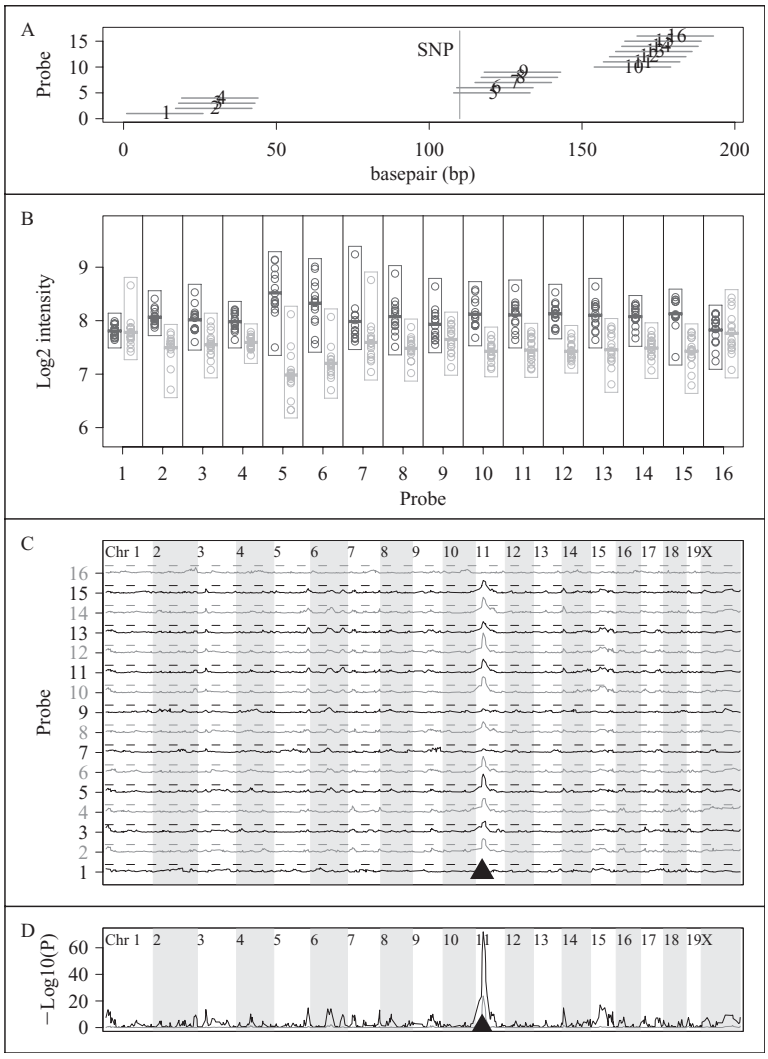

Figure 3: Genetic dissection of gene expression for short oligonucleotide arrays. Data are shown for gene 2410112O06Rik, probe set 93859_at, in a set of 30 RILs derived from a cross between parental strains B6 and D2. (A) Relative probe positions on the mRNA. The vertical line indicates a single nucleotide polymorphism (SNP) between strains B6 and D2. (B) Dissection of probe expression values. Per probe the expression values are split into two groups (dark and light grey) according to the type of allele at marker D11Mit19. Means per group are indicated by a horizontal stripe. The expression values are corrected for probe, batch and probe-specific batch effect. (C) QTL profiles on probe level for the 16 probes. The solid lines are the $-\log_{10}(P)$ values of the QTL effect. The dashed lines are the significance thresholds. The triangle indicates the position of the gene. (d) QTL profile on probe set level. The black line shows the significance of the QTL effect. The grey line shows the significance of the probe-specific QTL effect. The triangle indicates the position of the gene

measurements, $\gamma_{i\bullet\bullet}$ for the mean of the $n_{i\bullet}$ measurements from batch i, $\gamma_{\bullet j\bullet}$ for the mean of the $n_{\bullet j}$ measurements for QTL genotype j, and $\gamma_{ij\bullet}$ for the mean of all n_{ij} measurements in cell (i, j). The sums of squares in the ANOVA table (Table 3) are computed using the factor effects

$$B_i = \gamma_{i\bullet\bullet} - \gamma_{\bullet\bullet\bullet}$$

$$Q_j = \frac{1}{n_{\bullet j}}\left(n_{\bullet j}\,\gamma_{\bullet j\bullet} - \sum_{i=1}^{3} n_{ij} B_i\right) - \gamma_{\bullet\bullet\bullet}$$

Applying this model separately on each probe gives a QTL profile per probe

Table 3: Short oligonucleotide arrays: ANOVA table for probe data

Source of variation	Sum of squares	d.f.	*F*-value
Batch	$SS_{batch} = \sum_{i=1}^{3} n_{i\bullet} B_i^2$	2	
QTL effect	$SS_{QTL} = \sum_{i=1}^{3} n_{\bullet j} Q_j^2$	1	$\frac{26 \bullet SS_{QTL}}{SS_{error}}$
Error	$SS_{error} = SS_{total} - SS_{batch} - SS_{QTL}$	$n - 4 = 26$	
Total	$SS_{total} = \sum_{i=1}^{3} \sum_{j=1}^{2} \sum_{k=1}^{n_{ij}} (y_{ijk} - y_{\bullet\bullet\bullet})^2$	$n - 1 = 29$	

QTL analysis of probe set data

(Figure 3C). Most probes in this probe set appear to be *cis*-acting.

All probe expression values can also be modelled at once, obtaining a QTL profile at probe set level. The ANOVA model for QTL analysis on probe set level is

$$y_{ijkl} = m + P_k + B_i + PB_{ik} + Q_j + PQ_{jk} + M_{ijl} + e_{ijkl}$$

where y_{ijkl} is the log-intensity of the lth mouse ($l = 1, \ldots, n_{ijk}$) at probe k ($k = 1, \ldots, 16$), in batch i ($i = 1, 2, 3$) and QTL genotype j ($j = 1, 2$). In the model, P_k is the probe effect, B_i is the batch effect, PB_{ik} is the probe-specific batch effect, Q_j is the QTL effect, PQ_{jk} is the probe-specific QTL effect, M_{ijl} is the mouse error (note that we know that $\Sigma_{i,j} n_{ijk} = 30$) and e_{ijkl} is the probe-specific mouse error. We are not interested in the mean differences between probes, nor in the non-biological variation which is introduced by using arrays in three batches. By putting the factors probe, batch and probe-specific batch in the model, these unwanted sources of variation are removed from the data. A probe-specific QTL effect is included in the model in order to be able to detect cases in which the QTL effect varies over the probes in a probe set. A genome-scan was performed along all markers and marker D11Mit19 appeared to have the highest QTL effect. D11Mit19 is the nearest marker neighbour of the gene, indicating *cis*-activity. The expression values in Figure 3B are coloured according to the allele the corresponding mouse carries for this marker (light grey versus dark grey). The genome-wide QTL plot in Figure 3D is made by plotting the QTL effect for each marker. The figure also includes the probe-specific QTL effect. Notations such as $n_{i\bullet\bullet}$ and $y_{\bullet\bullet\bullet}$ etc are introduced as before. The factor effects, needed for the computation of the ANOVA table (see Table 4), are computed as follows:

$$B_i = y_{i\bullet\bullet\bullet} - y_{\bullet\bullet\bullet\bullet}$$

$$P_k = y_{\bullet\bullet k\bullet} - y_{\bullet\bullet\bullet\bullet}$$

$$BP_{ik} = y_{i\bullet k\bullet} - (y_{i\bullet\bullet\bullet} + B_i + P_k)$$

$$Q_j = \left[n_{\bullet j\bullet}^{-1} n_{\bullet j\bullet} y_{\bullet j\bullet\bullet} - \left(\sum_{i=1}^{3} n_{ij\bullet} B_i \right) \right] - y_{\bullet\bullet\bullet\bullet}$$

$$PQ_{jk} = n_{\bullet jk}^{-1} \left[n_{\bullet jk} y_{\bullet jk\bullet} - \left(\sum_{i=1}^{3} n_{ijk} B_i + P_k + \sum_{i=1}^{3} BP_{ik} \right) \right] - y_{\bullet\bullet\bullet\bullet}$$

and the fitted sum of squares can be computed via:

$$SS_{fitted} = SS_{batch} + SS_{probe} + SS_{b*p} + SS_{marker} + SS_{p*m} + SS_{mouse}$$

Table 4: Short oligonucleotide arrays: ANOVA table for probe set data

Source of variation	Sum of squares	d.f.	*F*-value
Batch	$SS_{batch} = \sum_{i=1}^{3} n_{i\bullet\bullet} B_i^2$	2	$\frac{SS_{batch}/2}{SS_{mouse}/26}$
Probe	$SS_{probe} = \sum_{k=1}^{16} n_{\bullet\bullet k} P_k^2$	15	$\frac{SS_{probe}/15}{SS_{error}/390}$
Batch * probe	$SS_{b*p} = \sum_{i=1}^{3}\sum_{k=1}^{16} n_{i\bullet k} BP_{ik}^2$	30	$\frac{SS_{b*p}/30}{SS_{error}/390}$
Marker	$SS_{marker} = \sum_{j=1}^{2} n_{\bullet j\bullet} Q_j^2$	1	$\frac{SS_{marker}}{SS_{mouse}/26}$
Probe * marker	$SS_{p*m} = \sum_{j=1}^{2}\sum_{k=1}^{16} n_{\bullet jk} PQ_{jk}^2$	15	$\frac{SS_{p*m}/15}{SS_{error}/390}$
Mouse	$SS_{mouse} = \sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{l=1}^{n_{ij\bullet}/16}\left(\sum_{k=1}^{16} y_{ijkl} - B_i - y_{\bullet\bullet\bullet\bullet}\right)^2 - SS_{marker}$	26	$\frac{SS_{mouse}/26}{SS_{error}/390}$
Error	$SS_{error} = SS_{total} - SS_{fitted}$	390	
Total	$SS_{total} = \sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{16}\sum_{l=1}^{n_{ijk}}(y_{ijkl} - y_{\bullet\bullet\bullet\bullet})^2$	479	

Note that the mouse error is taken as random factor.

Polymorphisms affect probe intensity

The significant probe-specific QTL effect is caused by probes 5 and 6 which show more differential expression than do the other probes (Figure 3B). This results in higher QTL peaks in Figure 3C for these probes. It is caused by a single nucleotide polymorphism between the B6 and D2 strains. On relative probe position 110 the sequence of strain B6 has a 'G' while D2 has a 'T'. Probes 5 and 6 have a 'G' on this position. The D2 mice hybridise less well because of this difference between the D2 mouse strain and the probes on the array.

An advantage of our modelling is that the probe-specific QTL effect traces probe sets in which probes respond differently, possibly caused by SNPs, splice variants, etc. This is not possible with the conventional summary methods,[8] since they average over the probes. For all probe sets that do not show a probe-specific QTL effect, it suffices to consider only the QTL effect of the analysis on probe set level.

This section showed how a genetic study of oligonucleotide arrays should be carried out, requiring again the use of non-standard analysis methods.

INFORMATION SYSTEMS FOR GENETICAL GENOMICS

The bioinformatics strategies presented above involve large data sets and complex sequences of statistical processing methods that would take months to execute by hand. Therefore, laboratories need tailored information infrastructures to support data handling, data processing and traceable research workflow. The well-known WebQTL internet service, specialised in well-curated genotype and phenotype data of mouse recombinant inbred lines, illustrates what such a tailored system can look like.[9,10] This system was initially developed just as a public repository of data, but an increasing amount of additional features have been added over time, urging a redesign of the system given the moving scope. The Molecular Genetics

Information System (MOLGENIS) has been implemented by the authors for microarray experiments on bacteria, plants, animals and clinical samples,[11] who are now in the process of completing QTL processing functionality in MOLGENIS4QTL.

Requirements for genetical genomics infrastructure

What criteria make an infrastructure such as WebQTL or MOLGENIS4QTL adequate? Five typical requirements for such molecular genetics information infrastructures are identified: (i) evolution ability to keep the system up to date in the fast-developing genomics field, (ii) a suitable data model that fits laboratory-specific conditions, (iii) suitable storage of data sets in the system, (iv) easy integration of processing software, and (v) low implementation and maintenance.[12] A suitable data model for QTL experiments includes marker maps, gene expressions sets, normalised expression sets, interval maps and graphics. These data model entities must be tailored to the specific research at hand because of variation in analytical methods (eg the differences between two-colour microarrays and oligonucleotide arrays), specificities of research topics (brain, ageing) and species (human, mouse, *Arabidopsis*, bacteria) and relevant parts of the wet-lab data management (such as array batches). Integrated processing methods include normalisation, interval mapping, visualisations but also more generic functionality such as searches, selections, task automation, cross-references to other resources and data import/export. MOLGENIS provides evolution ability using a code generator that takes a suitable data model as input and returns a low-maintenance web application. With this automated assembly process specialised variants of the bioinformatics infrastructure are produced in a short period and can be repeated for every new evolution of the research process.

DISCUSSION

In this paper the general concepts underlying genetical genomics have been outlined and illustrated. It has been shown that both two-colour cDNA arrays and short-oligonucleotide arrays in a genetical genomics set-up should not be designed and analysed following standard procedures for microarray analysis. For two-colour microarrays an optimal experimental design has been described in which individuals with contrasting marker fingerprints are compared at the same allele and a statistical model for QTL analysis has been provided. Alternative models and methods are described following Albers *et al.*[13] For oligonucleotide arrays it has been shown that the analysis differs from standard procedures by not taking the average signal over the probes and we have introduced statistical models for QTL mapping of probe set data.

The power of genetical genomics lies in its multifactorial efficiency. For example, in a RIL population of size 100, there are for each gene on average 50 individuals homozygous A and 50 B, giving a powerful *t*-test for each gene. In other words, good replication for each gene is automatically built-in in a genetical genomics experiment. Of course, expression profiling of 100 (or more) individuals is still a large job.

It has been demonstrated that by identifying *cis*-acting genes within a QTL region the expression data can be used to narrow down from QTL to gene. Furthermore, it has been shown how co-localisation of QTLs can uncover gene regulatory relationships. The integration of genome-wide QTL profiles has the potential of fully reconstructing gene regulatory networks. Genetical genomics can generate and combine such QTL profiles for any type of expression data (mRNA, protein, metabolite, quantitative phenotypes) and any type of organism.

Several genetical genomics studies have been performed on different organisms such as yeast,[3,14] maize,[15] mouse[10,15–18] and human.[15,18] However, the population sizes of these studies were relatively small. Larger populations are currently profiled in different projects. As soon as they are

available, the power of genetical genomics can be fully exploited and papers on the reconstruction of gene regulatory networks will appear. Information infrastructures such as WebQTL and MOLGENIS4QTL will provide the necessary working environments.

References

1. Jansen, R. C. and Nap, J. P. (2001), 'Genetical genomics: The added value from segregation', *Trends Genet*., Vol. 17(7), pp. 388–391.
2. Jansen, R. C. and Nap, J. P. (2004), 'Regulating gene expression: surprises still in store', *Trends Genet*., Vol. 20(5), pp. 223–225.
3. Yvert, G., Brem, R. B., Whittle, J. *et al*. (2003), 'Trans-acting regulatory variation in *Saccharomyces cerevisia*e and the role of transcription factors', *Nat. Genet*., Vol. 35(1), pp. 57–64.
4. Yang, Y. H., Dudoit, S., Luu, P. and Speed, T. P. (2001), 'Normalization for cDNA microarray data', in 'Microarrays: Optimal Technologies and informatics' , Bittner, M. L., Chen, X., Dorsel, A. N. and Dougherty, E. R. Eds, Proceedings of SPIE, vol. 4266, pp. 141–152.
5. Wolfinger, R. D., Gibson, G., Wolfinger, E. D. *et al*. (2001), 'Assessing gene significance from cDNA microarray expression data via mixed models', *J. Comput.Biol*., Vol. 8(6), pp. 625–637.
6. Fu, J. and Jansen, R. C. (2005), 'Optimal design and analysis of genetic studies on gene expression', unpublished MS, available from the authors.
7. Alberts, R., Bystrykh, L., de Haan, G. and Jansen, R. C. (2005), 'Genetics with oligonucleotide arrays', unpublished MS, available from the authors.
8. Irizarry, R. A., Bolstad, B. M., Collin, F. *et al*. (2003), 'Summaries of Affymetrix GeneChip probe level data', *Nucleic Acids Res*., Vol. 31 (4), p. e15.
9. Wang, J., Williams, R. W. and Manly, K. F. (2003), 'WebQTL: Web-based complex trait analysis', *Neuroinformatics*, Vol. 1 (4), pp. 299–308.
10. Chesler, E. J., Lu, L., Wang, J. *et al*. (2004), 'WebQTL: Rapid exploratory analysis of gene expression and genetic networks for brain and behavior', *Nat. Neurosci*., Vol. 7(5), pp. 485–486.
11. URL: http://www.molgenis.org
12. Swertz, M. A., De Brock, E. O., Van Hijum, S. A. *et al*. (2004), 'Molecular Genetics Information System (MOLGENIS): Alternatives in developing local experimental genomics databases', *Bioinformatics*, Vol. 20(13), pp. 2075–2083.
13. Albers, C. J., Jansen, R. C., Kok, J. *et al*. (2005), 'Simage: simulating microarray expression data', unpublished MS, available from the authors.
14. Brem, R. B., Yvert, G., Clinton, R. and Kruglyak, L. (2002), 'Genetic dissection of transcriptional regulation in budding yeast', *Science*, Vol. 296(5568), pp. 752–755.
15. Schadt, E. E., Monks, S. A., Drake, T. A. *et al*. (2003), 'Genetics of gene expression surveyed in maize, mouse and man', *Nature*, Vol. 422(6929), pp. 297–302.
16. Klose, J., Nock, C., Herrmann, M. *et al*. (2002), 'Genetic analysis of the mouse brain proteome', *Nat. Genet*., Vol. 30(4), pp. 385–393.
17. Bystrykh, L., Weersing, E., Dontje, B. *et al*. (2005), 'Uncovering regulatory pathways affecting hematopoietic stem cell function using ''genetical genomics''', *Nat. Genet*., Vol. 37(3), pp. 225–232.
18. Morley, M., Molony, C. M., Weber, T. M. *et al*. (2004), 'Genetic analysis of genome-wide variation in human gene expression', *Nature*, Vol. 430 (7001), pp. 743–747.